Transkript des Vortrags „Diversität und Diskriminierung durch KI“ von Laura Schelenz, Internationales Zentrum für Ethik in den Wissenschaften, Universität Tübingen, Stand: Mai 2024

Begrüßung

Guten Tag! Ich begrüße Sie herzlich zu diesem Vortrag zum Thema Diversität und Diskriminierung durch Künstliche Intelligenz. Mein Name ist Laura Schelenz und ich arbeite an der Universität Tübingen am Internationalen Zentrum für Ethik in den Wissenschaften. Ich beschäftige mich mit den sozialen Auswirkungen von Technik wie Künstlicher Intelligenz, also Computersystemen, die meist mit großen Datenmengen arbeiten und auf Automatisierung und Selbstlernen beruhen. Mich interessiert, welche moralischen und sozialen Vorstellungen, Werte und Ziele mit Technik verbunden sind und wie sich das in der Entwicklung der Technik ausdrückt. Also die Frage, wie kommen eigentlich gesellschaftliche Werte und Vorstellung in eine App oder ein System. Ich arbeite auch zum Feminismus, einer wissenschaftlichen und sozialen Bewegung, die darauf schaut, wie Frauen oder Minderheiten in der Gesellschaft und auch in Technik repräsentiert sind. Mich interessiert zum Beispiel, welche Auswirkungen Künstliche Intelligenz für Menschen hat, die normalerweise nicht als Norm wahrgenommen werden – dazu zählen Frauen, aber auch nicht-binäre Geschlechter, oder Menschen, die anderweitig als „divers“ wahrgenommen werden. Ich steige erstmal mit einer kleinen Geschichte ein, die verdeutlicht, was Technik wie Künstliche Intelligenz eigentlich mit Diversität und Diskriminierung zu tun hat.

Anfang des Jahres 2023 gab es eine große Begeisterung und gleichzeitig einen Aufschrei in der Technik-Welt. In nur kurzer Zeit ist die Nutzung des Chatbots ChatGPT rasant gestiegenen. Der Chatbot basiert auf Künstlicher Intelligenz und steht für die unglaubliche Leistungsfähigkeit von sogenannten Large Language Models (LLMs), also riesige Sprachmodelle, die Texte und mittlerweile auch Bilder konstruieren können.

Mit dem großen Erfolg von ChatGPT im Sinne der Nutzer:innenzahlen – nach nur zwei Monaten hatten bereits 100 Millionen Nutzer:innen den Chatbot ausprobiert – gab es auch einen Aufschrei in der Wirtschafts- und Technikelite. Ungewöhnlich deutlich haben führende Expert:innen vor den Risiken Künstlicher Intelligenz gewarnt. Prominente Beispiele sind der Nobelpreisträger für Wirtschaftswissenschaften Joseph E. Stiglitz und der Informatiker Geoffrey Hinton, einer der Mitbegründer der KI-Forschung. Während Stiglitz sich um die wirtschaftlichen Folgen von KI sowie wachsende Ungleichheit sorgt, warnt Hinton vor der Kontrolle von KI über die Menschheit.[1] Auch andere bekannte und einflussreiche Vertreter der Technikbranche haben sich kritisch geäußert, teilweise dystopische Bilder von der Übernahme der Menschheit durch sogenannte Künstliche allgemeine Intelligenz gemalt.

[1] Vgl. Bushwick, Sophie, „Unregulated AI Will Worsen Inequality, Warns Nobel-Winning Economist Joseph Stiglitz“, in: *Scientific American*, 01.08.2023. Online abrufbar unter: www.scientificamerican.com/article/unregulated-ai-will-worsen-inequality-warns-nobel-winning-economist-joseph-stiglitz/ [letzter Zugriff: 22.09.2023]; vgl. Taylor, Josh / Hern, Alex, „Godfather of AI’ Geoffrey Hinton Quits Google and Warns over Dangers of Misinformation“, in: *The Guardian*, 02.05.2023. Online abrufbar unter: www.theguardian.com/technology/2023/may/02/geoffrey-hinton-godfather-of-ai-quits-google-warns-dangers-of-machine-learning [letzter Zugriff: 02.10.2023].

Nun würde mich interessieren, ob Sie bereits einmal von dieser Frau gehört haben. Ihr Name ist Timnit Gebru.

Denn bereits im Jahr 2021 hatte die KI-Ethikerin Timnit Gebru zusammen mit Kolleg:innen in einem wissenschaftlichen Artikel auf die **Gefahren von LLMs** aufmerksam gemacht.[2] Gebru ist eine Schwarze Frau mit Wurzeln in Äthiopien und hatte bei Google eine Führungsrolle im Bereich Ethik inne. Sie wurde im Zusammenhang mit ihrem kritischen Artikel zu Large Language Models entlassen. Also bereits zwei Jahre bevor ChatGPT solch große Wellen schlug und sich die Eliten zu Wort meldeten, hatte Timnit Gebru vor KI-gestützten Sprachmodellen gewarnt. Anstatt jedoch besondere Aufmerksamkeit für ihre Kritik zu erlangen und öffentliche Diskurse zu prägen, wurde sie von Google entlassen.

Kritiker:innen von Google und deren Entscheidung, ihr zu kündigen, weisen in dieser Sache eine diskriminierende Praxis nach. Turner, Wood und D'Ignazio (2021) sprechen von einem sogenannten Playbook, also einer typischen Vorgehensweise im Hinblick auf den Umgang mit Schwarzen Frauen, Minderheiten und marginalisierten Geschlechtern, die sich mit Ungerechtigkeit in der Technikbranche befassen.[3] Hier geht es also um die Ermöglichung oder Verunmöglichung von Diversität in der Technikbranche selbst. Durch eine von weißen Menschen geprägte Unternehmenskultur, aber auch eine neoliberale Umgebung, wo Ethik und kritische Reflexion als Hindernis für Innovation und Profit gesehen werden, gibt es schwierige Bedingungen für Minderheiten. Ebenso geht es um die Diversität von Perspektiven auf künstliche Intelligenz, z.B. kritische Stimmen aus marginalisierten Gruppen, die selbst von Diskriminierung in der Technik-Industrie oder durch die KI-Anwendungen betroffen sind. Diese werden häufig nicht gehört oder aktiv unterdrückt, damit eben der Fortschritt der KI-Entwicklung nicht gehemmt wird. Wir dürfen jedoch nicht vergessen, das Beispiel hier bezieht sich auf den US-amerikanischen Kontext, auf das Silicon Valley. Dort gibt es wenig bis keine Regulierung von KI, anders als in Europa.

Okay, das nur zum Einstieg und Einstimmen. Das Thema Diversität und Diskriminierung im Zusammenhang mit Künstlicher Intelligenz ist also brandaktuell und gesellschaftlich ein besonders relevantes Thema. Worauf ich jedoch in dem Vortrag eingehen möchte, ist nicht so sehr die Diversität der Mitarbeiter:innen in Entwicklungs-Teams in der Technikindustrie. Das ist ein wichtiges Thema, wie wir gerade gesehen haben. Aber es gibt noch ein anderes wichtiges Thema im Zusammenhang mit Diversität und Diskriminierung, nämlich wie Diversitätskonzepte in die Technik selbst Einzug finden. Das hat zu tun mit dem Design einer KI-gestützten Anwendung, mit der Personalisierung von Services und mit der Klassifizierung und Überwachung von Nutzer:innen.

---

[2] Z. B. eine Verzerrung (auch Bias genannt) in den Datensätzen oder Klima- und Umweltfolgen des Betriebs riesiger Datenmengen; vgl. Bender, Emily M. / Gebru, Timnit / McMillan-Major, Angelina / Shmitchell, Shmargaret, „On the Dangers of Stochastic Parrots. Can Language Models Be Too Big?", in: Association for Computing Machinery (Hg.), *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency,* New York 2021, S. 612f.

[3] Vgl. Turner, Katlyn / Wood, Danielle / D'Ignazio, Catherine, „The Abuse and Misogynoir Playbook", in: *MIT Media Lab*, 01.01.2021. Online abrufbar unter: www.media.mit.edu/publications/abuse-and-misogynoir-playbook/ [letzter Zugriff: 02.07.2021].

Ich möchte in diesem Vortrag **drei Blöcke** behandeln. Der erste Block beschäftigt sich mit dem Verhältnis von Diversität und KI, also wie kann man Diversität eigentlich definieren, operationalisieren und in die KI-Anwendung einpflegen. Im zweiten Block zu den Vorteilen und Risiken von sogenannter „diversitätssensibler KI" klären wir dann, warum Informatiker:innen eigentlich Diversität nutzen, um ihre Technik zu optimieren. Welche Versprechen leiten sich daraus ab und welche Risiken gibt es dabei? Das Beispiel, welches ich für diese Diskussion mitgebracht habe, ist die KI-gestützte Gesichtserkennung.

Der dritte Block bietet einen Überblick über Ansätze zur kritischen Arbeit mit Diversität und KI. Hier geht es um Design, also wie können wir KI-Anwendungen entwickeln, sodass sie wirklich diversitätssensibel sind – und was ich mit „wirklich" meine erkläre ich dann später in diesem Block.

Teil 1: Das Verhältnis von Diversität und KI

Zunächst schauen wir auf das Verhältnis von Diversität und KI. Dafür müssen wir klären: Was ist Diversität überhaupt?

Diversität ist ein **mehrdeutiger Begriff**. Er kann sich auf die Beschreibung von Unterschieden (also einem deskriptivem Element) oder einen normativen Anspruch auf die Inklusion verschiedener Gruppen in die Gesellschaft beziehen. Als deskriptiver Begriff bezieht sich Diversität auf den "Unterschied" zwischen Menschen, Tieren, gesellschaftlichen Wertvorstellungen, Fähigkeiten von Mitarbeiter:innen usw. In der Biologie und Medizin bezieht sich Diversität auf unterschiedliche Arten oder Körperformen (Roughgarden, 2013). In den internationalen Beziehungen bezieht sich Diversität auf den Pluralismus der Werte, Normen und Interessen von Nationalstaaten (Williams, 2016). In der Betriebswirtschaft wird Diversität als unterschiedliche Fähigkeiten der Arbeitnehmer verstanden (Jack, 2016).

Als **normatives Konzept** hat Diversität einen Wert. Es geht also um Wertvorstellungen in Verbindung mit Vielfalt. Zum Beispiel meinen wir manchmal, wenn wir von Diversität sprechen, dass wir es gut finden, wenn unterschiedliche gesellschaftliche Gruppen Zugang zu den Ressourcen im Land haben. Es geht um Inklusion, Toleranz, Teilhabe, und Gerechtigkeit.

Sowohl deskriptive als auch normative Konzepte von Diversität sind **umstritten**. Insbesondere die Unterscheidung von Menschen, von Körperformen, geistigen oder körperlichen Fähigkeiten, Sprachkompetenzen, kulturelle und ethnische Herkunft, Geschlecht – all diese Unterscheidungen, die häufig ganz implizit angebracht werden in öffentlichen Diskursen, können hochproblematisch sein. Denn diese sozial konstruierten Unterschiede sind die Grundlage für Diskriminierung, für Rassismus und Sexismus oder die Ausgrenzung von Menschen mit Behinderung. Das ist ein wichtiger Punkt, der auch in Bezug auf KI später noch relevant wird.

Auch die normativen Werte, die wir mit Diversität assoziieren, sind umstritten! Erst mal klingt das alles super, Gleichberechtigung, Toleranz, Inklusion, Gerechtigkeit. Das wollen wir, oder? Aber hinter diesen Wertvorstellungen stecken ganz unterschiedliche Ansätze zur Behandlung gesellschaftlicher Gruppen. Gleichberechtigung ist ein gutes Beispiel. Natürlich sollen Menschen gleichbehandelt werden. Für eine Ungleichbehandlung brauchen wir gute Gründe – warum soll die eine Person einen Vorteil gegenüber einer anderen Person haben? Das finden wir unfair! Gleichzeitig ist es aber so, dass Gleichbehandlung nur in einer Gesellschaft gerecht ist, wo alle sozialen Gruppen dieselbe Ausgangslage haben. Nehmen wir das Beispiel Bildung.

Kinder mit Migrationshintergrund haben es häufig viel schwerer als deutsche Kinder, in der Schule und im Berufsleben aufzusteigen. Das kann viele Gründe haben – Sprachbarrieren, mangelnde finanzielle Ressourcen und Unterstützung der Eltern, vielleicht müssen die Eltern selbst erst die Sprache lernen, usw. – es gibt also nicht immer dieselbe Ausgangsituation für gesellschaftliche Gruppen. Hier fänden wir es gerecht, wenn die Gruppen, die eine schwierige Ausgangslage haben, einen Vorsprung, einen Vorteil erhalten, damit das Spielfeld wieder ausgeglichen ist. So könnte man argumentieren, dass Gleichberechtigung eigentlich nicht immer gut ist, sondern wir das Konzept der Gerechtigkeit mit Diversität in Verbindung bringen müssen.

Wie kommt jetzt die Diversität – entweder als Unterschied von „etwas“ oder als Wertvorstellung – in die KI?

Hier möchte ich zwei Phänomene beleuchten. Das erste die die Personalisierung. Netflix, Facebook, Amazon, Google und viele andere Dienste, die die meisten Menschen täglich benutzen, personalisieren die Inhalte, die wir sehen! Sie empfehlen Posts, Entertainment oder Produkte und arbeiten dabei mit speziellen Methoden aus der Informatik.

Eine dieser Methoden ist das collaborative filtering. Dies bedeutet, dass das KI-System Inhalte für eine bestimmte Nutzerin herausfiltert, die von anderen Nutzer:innen konsumiert werden, die der Nutzerin ähnlich sind. Also ähnliche Nutzer:innen sehen auch ähnliche Inhalte. Woher weiß das System, dass die Nutzerin ähnlich zu anderen ist und unter welchem Gesichtspunkt ähnlich? Bei collaborative filtering Methoden geht es ausschließlich um das Verhalten der Nutzer:innen. Also zwei sehr unterschiedliche Nutzer:innen (z.B. in Bezug auf Alter oder Bildung) können denselben Film mögen. Damit werden sie als ähnlich eingestuft und erhalten Empfehlungen für Filme, welche die jeweils andere Person auch konsumiert.

Und das sehen wir hier sehr schön bei Netflix. Auf der deutschen Webseite von Netflix wird die Personalisierung, die Netflix vornimmt, beschrieben. Hier steht, dass Netflix sich auf das Verhalten der Nutzerinnen konzentriert. Es werden also nicht explizit Diversitätskriterien wie Alter oder Geschlecht herangezogen, sondern es geht eher um das Verhalten und basierend darauf werden Kategorien von Nutzungsgruppen gebildet.

Aber diese Empfehlungen ausschließlich basierend auf Nutzungsverhalten können auch schwierig sein und zu Unzufriedenheit der Nutzer:innen führen. Nicht immer ist das Nutzungsverhalten ausschlaggebend für die eigenen Vorlieben. Vielleicht kauft man auf Amazon ein Geschenk für jemanden zu Weihnachten und ist selbst gar nicht an dem Produkt interessiert. Vielleicht teilt man einen Account mit dem Bruder oder der Schwester. Vielleicht nutzt man denselben Account einmal für die Freizeit und einmal für die Schule, das Studium oder die Arbeit. Informatiker:innen haben angefangen, das Nutzungsverhalten als nur einen von vielen wichtigen Aspekten bei der Personalisierung einzuschätzen. Immer wichtiger werden explizite Diversitätskriterien. Hier geht es dann darum, Gruppen von Nutzer:innen zu klassifizieren, z.B. nach Alter, Geschlecht oder Bildungsgrad, um die richtigen Nutzer:innen zu erreichen. Dies ist möglich über selfreports, also wenn Nutzer:innen ihre Daten selbst mitteilen, über Umfragen oder indem sie diese Informationen in einem Profil hinterlegen.

Also für Empfehlungen und Personalisierung wird die Diversität von Nutzer:innen auf unterschiedliche Art definiert und nutzbar gemacht, um die KI-Produkte zu verbessern und die Zufriedenheit der Nutzer:innen zu erhöhen.

Ein zweites Feld der Informatik, in dem Diversität eine große Rolle spielt, ist der Bereich **Machine Learning Fairness**. Dies ist ein Forschungsfeld der Informatik, welches Datenverzerrungen entgegenwirken und Diskriminierung durch KI vermeiden möchte. Hier spielt also der normative Aspekt von Diversität eine Rolle, während zuvor die Operationalisierung, also die Definition von deskriptiven Diversitätskategorien im Vordergrund stand.

Diskriminierung durch KI entsteht unter anderem durch verzerrte Datensätze. Die KI-Anwendungen werden mithilfe großer Datensätze trainiert, die wiederum die relevanten Informationen zu einem Thema bergen. Wenn zum Beispiel eine Gesichtserkennungs-Software entwickelt werden soll, wird dafür ein Datensatz mit Bildern von Gesichtern erstellt. Wenn eine Rekrutierungs-Software zur Vereinfachung von Bewerbungsprozessen entwickelt werden soll, dann könnten beispielsweise Lebensläufe oder Informationen von Bewerber:innen in solch einem Datensatz gesammelt sein. Häufig ist es leider so, dass diese **Datensätze verzerrt** sind. Das bedeutet, sie beinhalten besonders viele Informationen von einer Gruppe, zum Beispiel Gesichter von weißen Menschen, oder Informationen von männlichen Bewerbern. Das kann an der Datenverfügbarkeit liegen, aber auch andere Gründe haben. Im Ergebnis wird die KI-Anwendung dann automatisch so trainiert, dass sie besonders gut die Gesichter von Menschen mit heller Hautfarbe erkennt bzw. männliche Bewerber bevorzugt. So entsteht KI-gestützte Diskriminierung.

Machine Learning Fairness befasst sich mit den Definitionen und Methoden, die erforderlich sind, um diese Prozesse fair zu machen (Oneto & Chiappa, 2020). Dazu gehört das Erstellen neuer, fairer Datensätze sowie Methoden, mit denen Computermodelle so angepasst werden können, dass sie ein faires Ergebnis für unterschiedliche Bevölkerungsgruppen gewährleisten. Mit Machine Learning Fairness lassen sich auch unfaire Ergebnisse im Nachhinein korrigieren. Also, hier spielt Diversität also eine Rolle im Zusammenhang mit verzerrten Datensätzen und Diskriminierung.

Teil 2: Vorteile und Risiken von „diversitätssensibler“ KI

Bisher haben wir Theoretisches zum Thema Diversität und Methoden der Informatik gehört. Wir haben besprochen, dass es unterschiedliche Aspekte von Diversität gibt – einmal gibt es Klassifizierungen und einmal Wertvorstellungen. Wir haben auch besprochen, dass Diversität Einzug findet in das Design von KI-gestützten Produkten, wie etwa großen Plattformen oder Gesichtserkennung. Warum wird aber Diversität in die Technik eingebettet? Was sind die Vorteile davon?

Da sind wir nämlich schon beim zweiten Block zu den **Vorteilen und Risiken bei der Arbeit mit Diversität in der Technikentwicklung**. Ich habe hier geschrieben „diversitätssensible KI“ – das soll in unserem Kontext heute einfach bedeuten, dass es um eine KI-Anwendung geht, die explizit Diversität einbezieht.

Wir hatten bereits die Fälle Personalisierung und Machine Learning Fairness. Was sind die Vorteile davon?

Für Nutzer:innen ergeben sich durch Personalisierung Vorteile, da ihre persönlichen Vorlieben durch das KI-System bedient werden. Die Nutzer:innen erleben eine bessere Nutzung, also eine verbesserte **user experience**, was die Zufriedenheit der Nutzer:innen erhöht. Darüber hinaus gibt es Vorteile für die Unternehmen. Denn mehr Zufriedenheit bedeutet auch mehr Bindung

der Nutzer:innen an das Produkt und eine häufigere Nutzung. Also ergibt sich mehr **engagement**. Dadurch wird der Profit des Unternehmens erhöht. Bei dem Beispiel Machine Learning Fairness steht nicht die Zufriedenheit der Nutzer:innen im Vordergrund, sondern die Gleichberechtigung unterschiedlicher Betroffener der KI. Wenn wir jetzt an ein Rekrutierungssystem denken, das den Bewerbungsprozess in einem Unternehmen vereinfachen soll, dann sind natürlich die Nutzer:innen die Personalchefs und nicht die Bewerber:innen. Von der KI-Software betroffen sind allerdings die Bewerber:innen, sodass ihre Gleichberechtigung auf dem Spiel steht. Der Vorteil für die Bewerber:innen, wenn Diversität ganz bewusst einbezogen wird, ist es also, eine **faire Beurteilung** durch die KI-Anwendung zu erhalten. Es gibt aber auch Vorteile für die Technikentwickler:innen – durch fairness Methoden könnte die Technik schneller akzeptiert werden, einen besseren Ruf genießen und damit die Verbreitung der Technik beschleunigen und Profite für die Entwickler:innen erhöhen.

Es ist also wichtig, zu überlegen: wem nutzt der Rückgriff auf Diversität in der Technikentwicklung? Wer profitiert davon? Es sind nicht nur die Nutzer:innen oder Betroffenen. Die Unternehmen benutzen Diversität auch ganz strategisch, um ihre Gewinne zu vergrößern. Da geht es dann nicht um gesellschaftlichen Nutzen und Diversität im Sinne von Gerechtigkeit. Sondern es wird Diversität genutzt, um ökonomische Ziele zu erreichen.

Das bringt uns zu den Risiken.

Was sind Risiken, wenn Diversität in die Technik eingebettet wird? Ich möchte insbesondere auf einen Aspekt eingehen. Das ist der Aspekt der **Verschleierung von sozialen Kontexten und strukturellen Dynamiken**. Das wird später deutlich, was ich damit meine. Ich habe zur Diskussion der Risiken das Beispiel der KI-gestützten Gesichtserkennung mitgebracht.

Gesichtserkennungssoftware wird in vielen Bereichen eingesetzt, hauptsächlich um öffentliche Sicherheit herzustellen. Am Flughafen gibt es **Gesichtserkennungs-Technik**, wenn man durch die automatisierte Passkontrolle geht. In der Fußgängerzone und an öffentlichen Plätzen oder auch teilweise vor privaten Gebäuden wird die Technik in Verbindung mit Kamera-Überwachung genutzt – vor allem in Großbritannien und den USA wird Gesichtserkennungs-Software eingesetzt – in Deutschland vereinzelt, aber das kommt. Die Technik hilft Sicherheitsbehörden wie der Polizei, Straftäter:innen zu identifizieren und mit hinterlegten Bildern in der eigenen Datenbank abzugleichen. So können Kriminelle gefasst werden. Das Problem mit der Technik ist, dass sie nicht für alle Menschen gleich gut funktioniert und somit manche Menschen fälschlicherweise als verdächtig eingestuft werden und womöglich zu Unrecht von den Behörden verfolgt werden. Diese Problematik hat eine Wissenschaftlerin, damals am Massachsetts Institute of Technology in den USA adressiert, Joy Buolamwini, wir sehen sie hier auf der rechten Seite. Gemeinsam mit Timnit Gebru, der ehemaligen Ethikerin bei Google, die wir zu Beginn gesehen haben, hat sie eine Studie im Jahr 2018 herausgebracht. Die Studie zeigt, dass die Genauigkeit, mit der die weltweit führenden Gesichtserkennungs-Technologien arbeiten, für Menschen mit dunkler Hautfarbe und vor allem für Schwarze Frauen sehr viel geringer ist als für Menschen mit heller Haut. Symbolisch hält Joy Buolamwini hier eine weiße Maske hoch. Die Wissenschaftlerin hat die Technik nämlich an sich selbst getestet und bemerkt, dass die KI ihr eigenes Gesicht nicht erkennen konnte. Erst als sie eine weiße Maske über ihr Gesicht gestülpt hatte, wurde sie von der KI erkannt.

Was hat das nun mit Diversität und Diskriminierung zu tun? Es zeigt sich, dass KI-Anwendungen nicht gleich gut für verschiedene soziale Gruppen in der Gesellschaft

funktionieren. Im Falle der Gesichtserkennungs-Software kann das fatale Konsequenzen haben, denn Schwarze Menschen werden überproportional oft fehlidentifiziert und verhaftet. In den USA gibt es eine Geschichte der Diskriminierung von Schwarzen Menschen, die auf die Zeit der Sklaverei zurückgeht. Polizeigewalt und hohe Raten an Inhaftierung betreffen vor allem Afroamerikaner:innen oder Einwanderer mit dunkler Hautfarbe. Hier ergibt sich ein Bild, das auf eine Interaktion von technischer und sozialer Diskriminierung hinweist. Ich spreche jetzt vor allem im amerikanischen Kontext. Die Geschichte des Rassismus wird unbewusst durch KI fortgetragen und diskriminierende Muster werden verstärkt. Dies ist kein technisches Problem allein. Der soziale Kontext ist wichtig.

Okay, wie kann man nun das Problem der Fehlidentifikation von Menschen mit dunkler Hautfarbe und Frauen beheben?

Wir haben vorhin schon über Machine Learning Fairness gesprochen, die Methode in der Informatik, die gegen Verzerrungen und bias in Datensätzen und Algorithmen wirkt. Auch bei der Gesichtserkennung ist ein Problem, dass die Daten, mit denen gearbeitet wird, nicht ausgeglichen sind. Joy Buolamwini hatte in ihrer Studie herausgefunden, dass die **Daten für die Software**, die sie untersucht hat, vorrangig von weißen Männern stammten. Eine Lösung für dieses Problem ist es, einen ausgeglichenen, balancierten Datensatz mit Gesichtern von unterschiedlichen Gruppen zu erstellen.

Joy Buolamwini hat nun so einen Datensatz mit diversen Gesichtern erstellt und zwar hat sie dafür öffentliche Fotos von Abgeordneten aus Parlamenten in Europa und Afrika genutzt. Okay, das scheint erstmal einzuleuchten. Die Informationen von bislang diskriminierten Personen werden einbezogen und die Technik wird verbessert, sodass sie optimal die Gesichter von allen Menschen erkennen kann. Der Datensatz kann von Wissenschaftler:innen und Unternehmen auf der Webseite der Studie „gender shades“ nachgefragt werden. Er kann also tatsächlich genutzt werden, um Gesichtserkennungs-Systeme zu verbessern. Das ist ein ganz wichtiger Schritt in die richtige Richtung!

Es gibt jedoch zwei Probleme mit dieser Art von diversitätssensiblem Design. Das erste Problem ist, dass Diversitätsnarrative genutzt werden, um eine Technik zu legitimieren, die aber in ihrem Einsatz immer noch problematisch ist. Die Gesichtserkennung wird schließlich nach wie vor für die **Überwachung von Menschen** eingesetzt. Mit der diversitätssensiblen KI ist es nun aber so, dass die Überwachung noch viel effizienter funktioniert. Hier kann es dann sehr schnell zu **racial profiling** kommen, also der gezielten Überwachung von Schwarzen Menschen aufgrund der rassistischen Vorurteile gegen diese Gruppe. Gesichtserkennungs-Software wird zum Bespiel verstärkt dort eingesetzt, wo sich Minderheiten aufhalten, in der Innenstadt, in der Fußgängerzone, usw. Ich spreche hier immer noch im amerikanischen Kontext, weil in Deutschland das Thema Rassismus ein wenig anders gelagert ist.

Das zwei Problem, und das hängt eng mit dem ersten zusammen, ist, dass Diversität häufig losgelöst vom sozialen Kontext betrachtet wird. Diversität bezieht sich für Informatiker:innen meist auf technische Aspekte, Kategorien von Unterscheidung, aber der gesellschaftliche Hintergrund einer Problemlage wird verschleiert. Technische Maßnahme wie Machine Learning Fairness werden zu weit verbreiteten Lösungsansätzen für ein Problem, nämlich Diskriminierung, das eigentlich **soziale Ursprünge** hat. In der amerikanischen Geschichte können wir nämlich ganz deutlich sehen, dass es eine strukturelle Diskriminierung von Menschen mit dunkler Hautfarbe gibt. Ein Risiko bei der Arbeit mit Diversität in der

Technikentwicklung ist dann, dass auf Diversität zurückgegriffen wird, um ein Technikprodukt zu legitimieren und dadurch vielleicht die unterliegende soziale Diskriminierung (z.B. durch eine rassistische Überwachungs-Praxis) zu verschleiern. Ein KI-System könnte dann von der Öffentlichkeit oder Investoren als besonders gut wahrgenommen werden, wenn es als „diversitätssensibel" deklariert wird. Aber nur weil Diversität in einer Form in den Design-Prozess eingeflossen ist, heißt das nicht, dass die Technik automatisch sozial gerecht ist und für moralische Zwecke genutzt wird.

Teil 3: Critical Design

Dann kommen wir zum dritten Block. Beim dritten und letzten Block schauen wir auf Methoden und Herangehensweisen zur kritischen Arbeit mit Diversität in der Technikentwicklung.

Das Beispiel der Gesichtserkennung hat uns gezeigt, dass der soziale Kontext eine große Rolle spielt, wenn es um Gleichberechtigung und Gerechtigkeit im Zusammenhang mit einer KI-Anwendung geht. Ob die Technik Diversität der Nutzer:innen oder Betroffenen einbezieht, ist nicht allein relevant, sondern es zählt ebenso, in welchem Kontext die Technik verwendet wird. Hier kommen wir jetzt zu der Diskussion, was eine **„wirklich" diversitätssensible KI-Anwendung** ausmacht. Das ist eine Frage, mit der sich Wissenschaftlerinnen und Praktikerinnen im Bereich Critical Design auseinandersetzen.

Man könnte zum Beispiel sagen, dass KI erst dann „wirklich" diversitätssensibel ist, wenn die Informatiker:innen im Design-Prozess auch gesellschaftliche Ungleichheiten berücksichtigen. Das ist ein Ziel von Critical Design, in der Technikentwicklung Fragen von Machtverhältnissen und sozialen Unterschieden zu bearbeiten.

Critical Design unterscheidet sich zu mainstream oder klassischen Technikentwicklungsprozessen, indem soziale Gerechtigkeit im Fokus des Design-Prozesses steht. Rassismus, Sexismus, Diskriminierung von Menschen mit unterschiedlichen Voraussetzungen und Fähigkeiten, Altersdiskriminierung – all diese Phänomene spielen bei Critical Design eine Rolle, denn sie sollen auf keinen Fall durch die Technik verstärkt werden, sondern eher mithilfe von innovativen KI-Anwendungen abgebaut werden.

Critical Design betrachtet auch Machtbeziehungen zwischen Technikentwickler:innen und Nutzer:innen oder Betroffenen. Häufig ist es so, dass diejenigen, die von der Technik profitieren sollen oder die einen Anspruch auf Gleichbehandlung durch die KI haben, nicht in den Design-Prozess mit einbezogen werden. Critical Design versucht hier, den Bedürfnissen von marginalisierten und benachteiligten Nutzer:innen besondere Aufmerksamkeit zu geben. Eine Methode zur Inklusion von Minderheiten und der Öffentlichkeit, die normalerweise keinen Zugang zu Technikentwicklung haben, ist das **Partizipative Design**. Vertreter:innen des partizipativen Designs möchten KI-Entwicklung demokratisieren, sodass eine Vielzahl von Perspektiven, Wünschen und Bedürfnissen in der Technik abgebildet sein wird.

Eine kritische Haltung zu Design beinhaltet auch, Diversität im Kontext der gesellschaftlichen Machtbeziehung zu analysieren und zu bewerten. Machtstrukturen existieren oft als unbewusste, fortgeführte Diskriminierungen aus früheren Zeiten. Ein Blick in die Geschichte hilft hier, Bewusstsein für strukturelle Diskriminierung zu schaffen.

Zuletzt ist ein wichtiger Aspekt von Critical Design, die Rahmenbedingungen der Technikentwicklung selbst zu überprüfen. Denn die Voraussetzungen für das Design von Technik beeinflussen den gesamten Prozess und auch die Ziele des Designs. Das Silicon Valley in den USA ist natürlich ein gutes Beispiel für ein Umfeld, welches Fortschritt und Effizienz durch Technik fördern möchte. Häufig geht die Technikentwicklung mit technikdeterministischen Einstellungen einher. Das bedeutet, bei vielen Informatiker:innen herrscht die Überzeugung, dass Technik die Lösung für soziale Probleme ist und nur mithilfe von Technik gesellschaftliche Fortschritte erzielt werden können. Auch der Profit von Unternehmen steht im Silicon Valley im Vordergrund. Unter diesen Bedingungen kann es schwierig sein, eine KI-Anwendung kritisch zu analysieren und die Betroffenen bei der Weiterentwicklung der Technik einzubeziehen. Denn das kostet Zeit und führt womöglich zu monetären Einbußen bei den Unternehmen. Kritisches Design selbst findet daher selten in der klassischen Technikindustrie statt, sondern eher am Rande, in Nischen, an der Universität, in Vereinen oder Kollektiven.

Da wir über kritische Ansätze sprechen, möchte ich gerne **zuletzt noch auf die Arbeiten von Schwarzen Feminist:innen** eingehen. Der Schwarze Feminismus ist eine intellektuelle Tradition und eine Praxis des Aktivismus, bei dem die Belange Schwarzer Frauen in den Mittelpunkt gestellt werden. Die Theorie entstand in den USA und hat unter anderem das Konzept der **Intersektionalität** hervorgebracht. Es besagt, dass Diskriminierung nicht als isolierte Benachteiligung in einem Aspekt betrachtet werden sollte. Stattdessen geht es darum, wie verschiedene Strukturen der Benachteiligung zusammenspielen. Schwarze Frauen sind demnach nicht nur als Frauen benachteiligt und nicht nur als Menschen mit dunkler Hautfarbe. Sie sind an der Schnittstelle von Geschlecht und Hautfarbe benachteiligt, z.B. bei der Bewerbung für einen Job, für einen Studienplatz, bei der Vergabe von Sozialhilfe, usw.

Hier habe ich eine Übersicht mitgebracht von Frauen, die sich kritisch mit KI beschäftigen. Sie alle nehmen eine Schwarzfeministische Perspektive ein.

Safiya Noble hat 2018 das Buch „**Algorithms of Oppression**“ herausgebracht und beschreibt darin, wie Google einen racial bias aufweist. Noble hat untersucht, wie Schönheit und Weiblichkeit in Google dargestellt wird, zum Beispiel wenn man nach verschiedenen Schlagworten sucht. Die Suche nach „girls“ oder „women“ ergab zu der Zeit ihrer Recherche, also vor 2018, als Suchergebnis vorrangig sexualisierte Darstellungen von Mädchen und Frauen. Diese Darstellungen, vermittelt durch den Google Algorithmus, verstärken bestehende Stereotypen in der Gesellschaft. Ein Beispiel, welches in Safiya Nobles Studie vorkam, hat mit der unterschiedlichen Haartextur von weißen und farbigen Menschen zu tun. Eine Friseursalon-Inhaberin in the USA bietet Frisuren für Afroamerikanerinnen an. Sie möchte, dass ihr Salon auf Google Maps und im Internet gut zu finden ist. Doch wenn man bei Google nach „Black Hair“ sucht, findet man Bilder von Frisuren mit schwarzer Haarfarbe, nicht Frisuren mit der Haartextur von dunkelhäutigen Menschen. Dies ist ein bias, der darauf hinweist, dass die Belange und Interessen weißer Menschen durch Algorithmen bevorzugt behandelt werden.

Wir sehen auf dieser Folie auch noch Sasha Costanza-Chock. Costanza-Chock hat sich sehr intensiv mit der Diskriminierung von transgender-Personen durch KI beschäftigt und in im 2020 erschienen Buch „**Design Justice**“ das Beispiel des Body-Scanners am Flughafen analysiert. Der Bodyscanner am Flughafen basiert auf dem Abgleich von Daten in einer hinterlegten Datenbank mit den Daten (also den Körpermaßen) des Passagiers, der durch die

Sicherheitskontrolle am Flughafen geht. Wenn das System eine Anomalie findet, schlägt es Alarm. Bei Transgender Personen kann es sein, dass die Körperteile nicht mit der sichtlichen geschlechtlichen Assoziation übereinstimmen. Eine transgender Frau wird z.B. – auch richtigerweise - als Frau wahrgenommen und der Sicherheitsbeamte gibt dem System als Aufgabe, eine Frau nach möglichen versteckten Sicherheitsrisiken hin zu überprüfen. Nun schlägt aber das System im Intimbereich Alarm, weil es hier eine Anomalie feststellt. Umgekehrt kann es sein, dass Transgender Männer im Bereich der Brust laut System eine Anomalie aufweisen und dann in diesem Bereich von den Sicherheitsbeamten abgetastet werden. Das sind sehr erniedrigende und diskriminierende Erfahrungen, die zurückzuführen sind auf KI-Systeme, Mustererkennung und Klassifizierung, die nicht „wirklich" diversitätssensibel sind – nicht im Sinne der sozialen Gerechtigkeit.

Auf der Folie hier könnte man selbstverständlich noch Joy Buolamwini ergänzen und die Studie zur Gesichtserkennung. Denn obwohl sie Gesichtserkennung effizienter und wirkungsvoller machen wollte mit einem ausgeglichenen Datensatz und Fairness-Methoden, hat sie auch Aktivismus betrieben und auf die Gefahren der Gesichtserkennung und Überwachung hingewiesen. Also Joy Buolamwini hat durchaus technische und soziale Gedanken zur Gesichtserkennung in ihrer Arbeit miteinander verbunden.

Damit möchte ich gerne den Vortrag schließen.

Als Fazit lässt sich zusammenfassend wiederholen:

- Informatiker:innen und Entwickler:innen von Künstlicher Intelligenz verwenden Konzepte und Werte im Zusammenhang mit Diversität, um die eigenen Technik-Produkte zu optimieren!
- Methoden wie Personalisierung und Machine Learning Fairness sollen helfen, Nutzer:innen bessere Ergebnisse zu liefern und Diskriminierung durch KI zu reduzieren.
- Das Beispiel der Gesichtserkennung zeigt jedoch, dass der soziale Kontext einer KI-Anwendung wichtig ist, um das Diskriminierungspotenzial einer KI zu beurteilen.
- Technische Lösungen alleine reichen nicht – es braucht auch eine Auseinandersetzung mit sozialen Faktoren.
- Critical Design ist eine Methode, welche die Bedürfnisse marginalisierter Gruppen zentriert und Machtbeziehungen sowie Rahmenbedingungen der Technikentwicklung reflektiert.

Vielen Dank für Ihr Interesse und Ihre Aufmerksamkeit.

Kontakt:

Laura Schelenz

International Center for Ethics in the Sciences and Humanities (IZEW)

University of Tübingen

Wilhelmstraße 19

72074 Tübingen · Germany

Laura.schelenz@uni-tuebingen.de